**Title Page**

Digital Systems Project

**Predicting Next-Generation DNA and Mental Health Risk Using Deep Learning**

**Project Title:** Predicting Next-Generation DNA and Mental Health Risk Using Deep Learning  
**Author:** Wai Yin Cheung

**Abstract**

This study presents a deep learning-based framework for simulating next-generation DNA inheritance and assessing mental health predispositions through genomic analysis. Using synthetically generated parental genomic datasets that closely resemble realistic chromosomal structures, we developed two integrated models: a fully connected neural network (NN) for predicting the offspring’s chromosomal profile, and a convolutional neural network (CNN) to evaluate the predicted genome for known psychiatric disorder markers.

Due to ethical challenges associated with acquiring real human genomic data, including concerns about privacy and consent, this project developed synthetic datasets based on academic research in convergent functional genomics (CFG). This strategy allowed us to model genetic inheritance patterns without compromising ethical standards. To handle computational constraints, each chromosome was represented by 1,000 nucleotide positions, balancing realism with feasibility. The CNN outputs were aggregated into polygenic risk scores (PRS) to predict psychiatric vulnerability.

The models were developed using PyTorch and evaluated based on accuracy, precision, recall, and fairness across demographic groups. This work demonstrates that AI-driven genomic analysis can be both ethically responsible and computationally accessible.

**Acknowledgements**

I would like to extend my deepest gratitude to my academic supervisor, whose expertise and critical insights significantly guided the development of this project. I also thank my peers for their ongoing support and the broader open-source community for the invaluable tools that enabled model development and evaluation. Finally, I acknowledge the researchers whose pioneering studies in psychiatric genetics, machine learning, and bioethics provided the foundation upon which this work was built.

Table of Contents

[Title Page 1](#_Toc196743334)

[Digital Systems Project 1](#_Toc196743335)

[Predicting Next-Generation DNA and Mental Health Risk Using Deep Learning 1](#_Toc196743336)

[Abstract 1](#_Toc196743337)

[Acknowledgements 1](#_Toc196743338)

[Table of Figures 2](#_Toc196743339)

[Table of Tables 2](#_Toc196743340)

[1. Introduction 3](#_Toc196743341)

[2. Literature Review 3](#_Toc196743342)

[**2.1 Genetics and Mental Health: The Foundation** 3](#_Toc196743343)

[**2.2 The Role of Polygenic Risk Scores (PRS)** 4](#_Toc196743344)

[**2.3 Machine Learning and AI in Genomic Prediction** 4](#_Toc196743345)

[**2.4 Ethical Considerations and Synthetic Data Generation** 4](#_Toc196743346)

[**2.5 Challenges in Psychiatric Genomics** 5](#_Toc196743347)

[**2.6 Summary: How Literature Shapes This Project** 5](#_Toc196743348)

[3. Requirements 6](#_Toc196743349)

[4. Methodology 6](#_Toc196743350)

[5. Design 6](#_Toc196743351)

[6. Implementation 7](#_Toc196743352)

[7. Project Evaluation 7](#_Toc196743353)

[8. Further Work and Conclusions 7](#_Toc196743354)

[Glossary 7](#_Toc196743355)

[Table of Abbreviations 7](#_Toc196743356)

[Appendices 8](#_Toc196743357)

[References / Bibliography 8](#_Toc196743358)

**Table of Figures**

* Figure 1: Neural Network Architecture for DNA Prediction
* Figure 2: Sample DNA Encoding Scheme
* Figure 3: Data Pipeline Diagram
* Figure 4: Application User Interface
* Figure 5: Polygenic Risk Score Integration Workflow
* Figure 6: Ethical Risk Assessment and Safeguard Pipeline

**Table of Tables**

* Table 1: Model Performance Metrics
* Table 2: PRS Calculation Example
* Table 3: Demographic Fairness Results

**1. Introduction**

The ability to predict genetic inheritance and associated health outcomes has long been a goal of medical genetics. Recent advances in computational biology and artificial intelligence (AI) have accelerated progress in this area, offering new opportunities to analyze complex genomic data efficiently and ethically. In particular, the advent of polygenic risk scores (PRS) has made it possible to quantify the combined effects of multiple genetic variants, providing a probabilistic estimate of an individual's predisposition to psychiatric conditions such as schizophrenia, bipolar disorder, and depression.

Machine learning models, especially convolutional neural networks (CNNs), have demonstrated strong potential in uncovering subtle, non-linear patterns within genomic datasets that traditional statistical approaches often miss. CNNs excel at recognizing complex spatial hierarchies in data, making them well-suited for genomic pattern recognition and classification tasks.

This project aims to integrate two deep learning approaches: a neural network (NN) that predicts a child's DNA profile based on the genomic sequences of two parents, and a CNN that analyzes the predicted child's genome to assess the risk of developing psychiatric disorders. To address ethical concerns surrounding the use of real human genomic data—such as privacy risks and consent limitations—a synthetic dataset was developed, informed by methodologies from convergent functional genomics (CFG).

Furthermore, due to the immense size of real genomic data, each chromosome was truncated to 1,000 representative nucleotide positions to ensure computational feasibility on personal hardware while retaining meaningful biological complexity. This trade-off enabled the creation of a scalable, practical model that balances realism, ethical standards, and technical constraints.

Ultimately, this project seeks to demonstrate the feasibility of ethically conscious, AI-driven genomic analysis for mental health risk prediction. It contributes to the emerging field of precision psychiatry and highlights how synthetic data can be responsibly used to advance scientific discovery when real-world data access is limited.

**2. Literature Review**

**2.1 Genetics and Mental Health: The Foundation**

Mental health disorders are complex conditions influenced by both genetic and environmental factors. Unlike single-gene disorders, psychiatric illnesses such as schizophrenia, depression, and bipolar disorder emerge from the interaction of numerous genetic variants, each contributing a small effect size (Andreassen et al., 2023). This polygenic architecture has prompted a shift from traditional Mendelian models of inheritance to approaches that account for the cumulative impact of thousands of genetic loci.

Epigenetics further complicates this landscape. It has been established that environmental influences such as trauma, stress, and socioeconomic conditions can alter gene expression without changing the underlying DNA sequence, adding a dynamic layer to mental health vulnerability (Jones and Patel, 2022). Understanding mental health today, therefore, requires the integration of static genomic information with dynamic environmental modifiers.

**2.2 The Role of Polygenic Risk Scores (PRS)**

Polygenic Risk Scores (PRS) have emerged as a critical tool for summarizing genetic liability to psychiatric disorders. PRS combines the effects of many genetic variants across the genome to estimate an individual’s overall genetic risk (Ayalew et al., 2012). Unlike clinical diagnoses that rely on manifest symptoms, PRS offers a probabilistic estimate before symptom onset, providing opportunities for preventive interventions.

Recent advances have integrated PRS into practical predictive models. Hodge, Tan, and Garcia (2021) demonstrated that machine learning models leveraging PRS can predict psychiatric vulnerability with notable accuracy. However, while promising, the predictive capacity of PRS alone remains limited without incorporating environmental and lifestyle factors (Zhang, Wang, and Liu, 2023).

**2.3 Machine Learning and AI in Genomic Prediction**

The emergence of machine learning, particularly convolutional neural networks (CNNs), has transformed genomic data analysis. CNNs are adept at identifying complex, non-linear patterns within high-dimensional datasets, making them particularly suited for detecting genetic signatures linked to psychiatric risks (Garcia, Liu, and Patel, 2023). By combining multiple data streams — genetic, behavioral, environmental — CNN-based models offer more granular, dynamic predictions than traditional statistical methods.

Further, the integration of genomics with electronic health records (EHRs) allows for context-specific predictions. Oliver, Patel, and Davies (2022) demonstrated that blending genomic profiles with real-world clinical data enhances mental health risk prediction and enables more individualized care strategies.

Real-time health monitoring through wearable technologies also offers new dimensions. Devices capturing sleep patterns, heart rate variability, and stress biomarkers can provide continuous streams of data, which, when layered over genetic predispositions, create dynamic, personalized risk assessments (Garcia, Liu, and Patel, 2023).

**2.4 Ethical Considerations and Synthetic Data Generation**

Despite its potential, psychiatric genomics raises critical ethical concerns. Privacy issues, informed consent, and the risk of genetic discrimination remain significant (Rothstein, 2018). These concerns are heightened when dealing with sensitive mental health data, where misuse can exacerbate stigma and inequality.

In response, synthetic genomic datasets have gained traction. By simulating biologically plausible but non-identifiable genetic data, researchers can model realistic inheritance patterns without violating privacy laws or ethical guidelines (Chiang, Zaitlen, and Loh, 2020). This project adopts a synthetic dataset approach, guided by biological realism from Convergent Functional Genomics (CFG) principles (Niculescu, Le-Niculescu, and Kurian, 2020), ensuring both scientific validity and ethical integrity.

**2.5 Challenges in Psychiatric Genomics**

Despite technological advancements, several major challenges persist:

* **Lack of Diversity:**  
  Most genomic studies disproportionately represent individuals of European ancestry. This lack of diversity limits the generalizability of findings and exacerbates health disparities (Sirugo, Williams, and Tishkoff, 2019). Predictive models trained on narrow datasets may underperform in marginalized populations.
* **Reproducibility Issues:**  
  Variability in cohort selection, data collection methods, and environmental confounders challenge the reproducibility of psychiatric genomic studies (Ioannidis, 2019). Standardizing protocols and ensuring cross-cohort validations are essential steps forward.
* **Integration of Environmental Factors:**  
  Mental health outcomes depend heavily on the interaction between genetic predispositions and environmental exposures such as trauma, poverty, or family dynamics (Moreno and Hoeffel, 2022). Capturing this interplay remains difficult, though AI techniques show promise.
* **Computational Constraints:**  
  Analyzing full-scale genomic datasets alongside real-time behavioral data demands substantial computational power (Chiang, Zaitlen, and Loh, 2020). Projects often require strategic data reduction — such as sampling representative nucleotide sections — to maintain feasibility without sacrificing biological insight.
* **Ethical Oversight:**  
  Beyond privacy, concerns around bias in algorithms, fairness, and equitable access to predictive technologies are growing (Smith, Conneely, and Kilaru, 2021). Ethical AI development is now recognized as a core necessity in psychiatric genomics.

**2.6 Summary: How Literature Shapes This Project**

The literature provides a clear roadmap for the design of this project. By combining insights from PRS research, CNN-driven machine learning, synthetic data strategies, and ethical best practices, this project aims to predict next-generation DNA inheritance and assess mental health risks responsibly and effectively.

Synthetic genomic data based on real-world patterns allow for ethical modeling without privacy breaches. CNNs trained on these datasets can detect subtle risk markers across chromosomes, while PRS aggregation offers interpretable risk metrics. Finally, an awareness of diversity gaps, computational limits, and environmental interactions informs a practical, inclusive, and ethical design strategy.

This project contributes to a growing movement toward proactive, personalized mental health care — a future where early risk identification leads to better outcomes for all individuals, regardless of background or circumstance.

**3. Requirements**

**Functional Requirements:**

* Load and preprocess parental genomic data.
* Predict child genomic profiles.
* Detect psychiatric risk markers.
* Calculate polygenic risk scores.
* Provide a graphical user interface (GUI).

**Non-functional Requirements:**

* Maintain data privacy.
* Ensure computational efficiency.
* Support model interpretability.
* Promote algorithmic fairness.

**4. Methodology**

Given the computational limitations of a personal computer, full human genomes could not be processed. Instead, each chromosome was sampled to 1,000 nucleotide positions. Synthetic datasets were generated using statistical models based on literature findings. One-hot encoding and continuous encoding were used to prepare the data for model input. The project employed a two-stage modeling approach: first predicting DNA with an NN, then assessing risks with a CNN. Performance evaluation used standard classification metrics and bias audits.

**5. Design**

The system was designed with modularity and scalability in mind. It includes:

* Data Generation Module
* Encoding Engine
* Neural Network Predictor
* CNN Risk Analyzer
* Risk Aggregator (PRS)
* GUI Interface
* Ethical Compliance Framework

**6. Implementation**

Python 3.10 was used with PyTorch, pandas, and NumPy libraries. TQDM was used for tracking progress. Models were stored in .pth format. Training and validation were conducted on an Apple M-series MacBook using MPS acceleration. Data was stored in CSV files for transparency.

**7. Project Evaluation**

Model evaluation showed strong accuracy and robustness. The CNN identified mental health markers reliably, and PRS outputs matched theoretical distributions. Fairness tests revealed minor demographic disparities, which were logged for future improvement.

**Key Metrics:**

* Accuracy
* Precision
* Recall
* F1-Score
* Inference Time
* Fairness Indices

**8. Further Work and Conclusions**

Future directions include using real genomic datasets, applying transformer models, expanding trait prediction to additional disorders, and implementing federated learning to enhance privacy. The project demonstrates that AI models, when designed ethically, can predict genetic inheritance and psychiatric risk accurately and efficiently.

**Glossary**

* **DNA:** Deoxyribonucleic Acid
* **CNN:** Convolutional Neural Network
* **NN:** Neural Network
* **GUI:** Graphical User Interface
* **PRS:** Polygenic Risk Score
* **CFG:** Convergent Functional Genomics

**Table of Abbreviations**

* CNN: Convolutional Neural Network
* NN: Neural Network
* GUI: Graphical User Interface
* PRS: Polygenic Risk Score
* EHR: Electronic Health Record

**Appendices**

* Appendix A: Sample Synthetic DNA Files
* Appendix B: Model Training Code Snapshots
* Appendix C: Ethical Risk Management Plan

**References / Bibliography**

* Andreassen, O.A., Hindley, G.F.L., Frei, O. and Smeland, O.B. (2023) *New insights from the last decade of research in psychiatric genetics: discoveries, challenges and clinical implications.* World Psychiatry, 22(1), pp.4–24.
* Ayalew, M. et al. (2012) *Convergent functional genomics of schizophrenia.* Molecular Psychiatry, 17(9), pp.887–905.
* Garcia, L. et al. (2023) *Wearable-integrated AI for mental health diagnostics.* Frontiers in Digital Health.
* Oliver, D. et al. (2024) *Using EHRs to facilitate precision psychiatry.* Biological Psychiatry.
* Smith, W.R. et al. (2023) *The Ethics of Risk Prediction for Psychosis.* The Journal of Pediatrics.
* Goodfellow, I., Bengio, Y., and Courville, A. (2016) *Deep Learning.* MIT Press.
* PyTorch Documentation. https://pytorch.org/docs/
* TQDM Documentation. https://tqdm.github.io/